P-2209

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-187753
(43)Date of publication of application : 21.07.1998

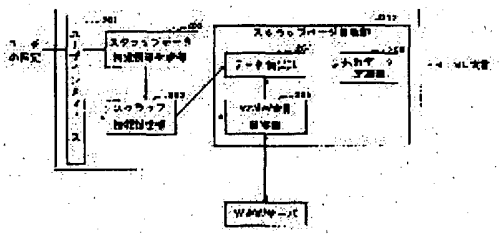
(51)Int.CI. G06F 17/30
G06F 12/00
G06F 12/00

(21)Application number : 08-356218
(22)Date of filing : 25.12.1996

(71)Applicant : NEC CORP
(72)Inventor : SUGIURA ATSUSHI

(54) WWW INFORMATION EXTRACTION SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide an information required by a user on a WWW with a little burden by generating and storing information for specifying the start and end position of designated data.
SOLUTION: The start and end positions of data required for a user himself are designated in a WWW document by a user interface 201, and the designated contents are specified by a scrap data specification information part 202. The pair of URL and scrap data specification information of a document, for which a user designates scrap data, is stored in a scrap information storage part 203. The latest WWW document corresponding to the stored URL is possessed by a possessing part 205 and based on the stored scrap data specification information, data required for the user are extracted out of the newly possessed WWW document by a data extracting part 204. Afterwards, the extracted data are collected into one page by a link part 206.



LEGAL STATUS

[Date of request for examination]	25.12.1996
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	2867986
[Date of registration]	25.12.1998
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平10－187753

(43)公開日　平成10年(1998)７月21日

(51)Int.Cl.[6]		識別記号	庁内整理番号	Ｆ１			技術表示箇所
Ｇ０６Ｆ	17/30			Ｇ０６Ｆ	15/403	３３０Ｚ	
	12/00	５１５			12/00	５１５Ａ	
		５４７				５４７Ｈ	
					15/40	３１０Ｆ	
					15/401	３４０Ａ	

審査請求　有　　請求項の数５　ＦＤ　（全９頁）

(21)出願番号　特願平8－356218

(22)出願日　平成８年(1996)12月25日

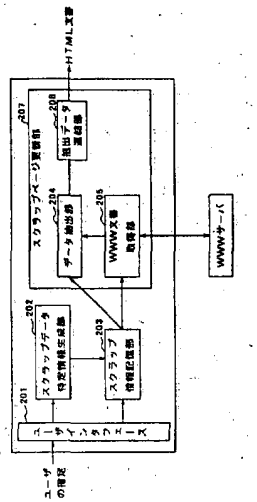
(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　杉浦　淳
東京都港区芝五丁目７番１号　日本電気株式会社内

(74)代理人　弁理士　加藤　朝道

(54)【発明の名称】　WWW情報抽出システム

(57)【要約】
【課題】日々更新されるWWW文書から必要な情報のみを効率よく少ない手間で抽出するWWW抽出システムの提供。
【解決手段】先ずユーザがWWW文書中で必要とするデータが存在する箇所を指定し、その指定箇所を特定するための情報を生成・保存しておき、ここで生成した情報を利用して、更新されたWWW文書からユーザが必要とするデータのみを抽出する。



【特許請求の範囲】
【請求項１】WWW（World Wide Web：ワールドワイドウェブ）サイトから情報を取得するWWW情報抽出システムにおいて、
ユーザがWWW文　について必要とするデータの存在する箇所を指定した際に、前記WWW文書における前記指定箇所を特定するための情報（「特定情報」という）を自動生成するとともに、生成された該特定情報を保存しておき、これ以降、新たに取得したWWW文書から、前記特定情報を利用して、前記ユーザが必要とするデータのみを抽出する、ことを特徴とするWWW情報抽出システム。
【請求項２】前記指定箇所を特定するための特定情報が、前記WWW文　において前記ユーザが指定したデータの開始および終了箇所を指示する情報からなり、該特定情報を、前記WWW文　のＵＲＬ（Uniform　Resource　Locator：ユニフォームリソースロケータ）と対応させて保存する、ことを特徴とする請求項１記載のWWW情報抽出システム。
【請求項３】保存されている前記特定情報に基づき、新たに取得したWWW文　中について、前記ユーザが先に指定したデータに対応するデータの開始箇所と終了箇所を判断し、前記開始箇所と前記終了箇所の間にあるデータを抽出することにより、前記ユーザが必要としていると考えられるデータを前記新たに取得したWWW文書中から抽出する、ことを特徴とする請求項２記載のWWW情報抽出システム。
【請求項４】保存されている全てのＵＲＬと前記特定情報の組に対して新たに取得したWWW文書の開始箇所と終了箇所の間にあるデータを抽出し、前記新たに取得したWWW文書が複数ある場合、各WWW文書から抽出したデータを一つの文書にまとめて提示する手段を備えた、ことを特徴とする請求項２記載のWWW情報抽出システム。
【請求項５】WWW（World　Wide　Web：ワールドワイドウェブ）サイトから情報を取得するWWW情報抽出システムにおいて、
ユーザがWWW文書中のデータの一部を指定することを可能とする手段と、
ユーザが指定したデータの開始および終了箇所を特定するための情報を生成するデータ特定情報生成手段と、
ユーザがデータを指定したWWW文書のＵＲＬ（Uniform　Resource　Locator）および前記データ特定情報生成手段で生成された情報を記憶する記憶手段と、
前記WWWサイトからWWW文書を取得するWWW文書取得手段と、
前記データ特定情報生成手段で生成され前記記憶手段に記憶されている情報を用いてユーザが先に指定したデータに対応するデータを前記WWW文　から抽出する手段と、
を備え、
WWW文書の内容が更新されたとしても、ユーザが指定したデータに対応するデータを抽出する、ことを特徴とするWWW情報抽出システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、WWW（World Wide Web）サイトからWWW文書を取得するシステムに関し、特にユーザが必要とする情報のみを抽出することを可能とするWWW情報抽出システムに関する。
【０００２】
【従来の技術】WWWサイトに存在する情報を取得し閲覧するためのツールとして、WWWブラウザがある。ユーザは目的のWWW文書のＵＲＬ（Uniform　Resource　Locator：インターネットやWWWでファイルを識別するために使われる規格）を指定することにより目的の情報を閲覧することが可能である。
【０００３】また、ＵＲＬ指定のためのユーザの手間を簡略化する目的で、オートパイロットツールが開発されている。これを利用すると、ユーザが指定した日時や時間間隔で、ユーザが予め指定しておいた複数のＵＲＬのWWW文書をまとめて取得することが可能である。
【０００４】
【発明が解決しようとする課題】WWWブラウザを用いてWWWサイトに存在するWWW文書から必要な情報を得るためには、ユーザは、目的のWWW文書のＵＲＬを指定し、その文書がWWWブラウザ上に表示された後に、文書をスクロールしながら目視により検索したり（人手による検索）、あるいは文字列検索機能を利用するといった作業を行なう必要がある。
【０００５】そして、複数のWWW文書から情報を取得する場合には、上記の作業を繰り返し行なうことになるが、これらのWWW文書が日々更新されるものである場合には、かなりの頻度で、上記と同様の作業を繰り返し行なわなければならない、ことになる。このように、WWW文書が日々更新される場合、複数のWWW文書から情報を得るための操作が煩雑なものとなって、ユーザの負担が増大し、この検索作業のために、所望する情報を迅速に得ることが困難となる。
【０００６】また、上記したオートパイロットツールを用いることにより、複数のWWW文書を一括したローカルマシン上へダウンロードすることは可能であるが、この場合でも、ユーザはローカルマシン上にあるファイルを自分で開き、必要な情報を探さなければならない。
【０００７】このため、ユーザが必要な情報を閲覧するための手間は、WWWブラウザを用いる場合と、それほど変わりはない。
【０００８】したがって、本発明は、上記した事情に鑑みてなされたものであって、その目的は、上記従来方式の問題点を解消し、WWW上でユーザが必要とする情報を少ない負担で得ることを可能とする、WWW情報抽出システムを提供することにある。

【０００９】
【課題を解決するための手段】前記目的を達成するため、本発明のWWW情報抽出システムは、ユーザが必要とする情報のみを複数のWWWページから自動的にスクラップし、１つの文書にまとめてユーザに提示する、ようにしたものである。
【００１０】すなわち、本発明のWWW情報抽出システムは、WWW（World　Wide　Web：ワールドワイドウェブ）サイトから情報を取得するシステムにおいて、ユーザがWWW文書中のデータの一部を指定することを可能とする手段と、ユーザが指定したデータの開始および終了箇所を特定するための情報を生成するデータ特定情報生成手段と、ユーザがデータを指定したWWW文書のＵＲＬ（UniformResource　Locator）および前記データ特定情報生成手段で生成された情報を記憶する記憶手段と、前記WWWサイトからWWW文書を取得するWWW文書取得手段と、前記データ特定情報生成手段で生成され前記記憶手段に記憶されている情報を用いてＭユーザが先に指定したデータに対応するデータをWWW文書から抽出する手段と、を備え、WWW文書の内容が更新されたとしても、ユーザが指定したデータに対応するデータを抽出する、ことを特徴としたものである。
【００１１】
【発明の実施の形態】本発明の実施の形態について以下に説明する。本発明は、その好ましい実施の形態において、（ａ）ユーザがWWW文書中で自分が必要とするデータの開始箇所と終了箇所を指定することを可能とするユーザインタフェース（図２の２０１）、（ｂ）ユーザが前記ユーザインタフェース上で指定したデータ（以下「スクラップデータ」という）の開始および終了箇所を特定するためのスクラップデータ特定情報を生成するための手段（図２の２０２）、（ｃ）ユーザがスクラップデータを指定した文書のＵＲＬおよびスクラップデータ特定情報の組（以下「スクラップ情報」という）を記憶するためのスクラップ情報記憶部（図２の２０３）を有する。
【００１２】ユーザは複数のWWWページの複数箇所に対してスクラップデータの指定を行うことを可能とする。
【００１３】システムは、ユーザが指定したそれぞれのスクラップデータに対して、スクラップデータ特定情報を生成し、スクラップ情報記憶部（図２の２０３）に記憶する。
【００１４】本発明は、その好ましい実施の形態において、さらに、上記（ａ）～（ｃ）のに加え、（ｄ）スクラップ情報記憶部（図２の２０３）に記憶されたＵＲＬに対応する最新のWWW文　を取得する手段（図２の２０５）、（ｅ）スクラップ情報記憶部（図２の２０３）に記憶されたスクラップデータ特定情報に基づき、新たに取得したWWW文書中から、ユーザが必要とするデータの開始箇所と終了箇所を特定し、開始箇所と終了箇所の間にあるデータを抽出する手段（図２の２０４）、（ｆ）スクラップ情報記憶部（図２の２０３）に記憶された全てのＵＲＬとスクラップデータ特定情報の組に対して、上記（ｄ）および（ｅ）における処理を行った後に、抽出したデータを１つのページにまとめる手段（図２の２０６）、を備えて構成されている。
【００１５】本発明は、その好ましい実施の形態において、ユーザは、最初にWWW文書において自分が必要とするデータの開始箇所と終了箇所を指定しておけば、以後はシステム側で、最新のWWW文書を取得し、新たに取得した文書からユーザが必要としていると考えられるデータのみを自動抽出する。
【００１６】このため、新たに取得したWWW文書中からユーザ自らが自分の必要とする情報を検索する必要はない。
【００１７】また、ユーザが必要とする情報が複数のWWW文書に存在する場合であっても、本発明の実施の形態に係るWWW情報抽出システムは、ユーザが必要とするデータを、各WWW文　から抽出し、抽出したデータを１つの文書にまとめてユーザに提示するため、ユーザは各WWW文書を１つ１つ開いて内容を閲覧する必要はなく、自分の必要な情報のみを一括して閲覧することが可能である。
【００１８】このため、WWWサイトから必要な情報を取得し検索するために要する作業コストを大幅に軽減することが可能である。
【００１９】
【実施例】上記した本発明の実施の形態について更に詳細に説明すべく、本発明の実施例について以下に説明する。
【００２０】図１は、本発明を一実施例のシステムの全体構成の概略をブロック図にて示したものである。図１を参照すると、本実施例に係るシステムは、データを表示するためのディスプレイ装置１０１と、マウスなどのポインティングデバイスおよびキーボードといった入力装置１０２と、スクラップデータ特定情報などを記憶するための記憶装置１０３と、これらディスプレイ装置１０１、入力装置１０２、及び記憶装置１０３を制御して処理を実行するコンピュータ１０４と、を備えて構成される。
【００２１】図２は、本発明の一実施例のWWW情報抽出システムの構成を機能ブロック図にて示したものである。
【００２２】図２を参照すると、WWW情報抽出システムは、ユーザがWWW文　の特定箇所を指定するためのユーザインタフェース２０１と、ユーザが指定したデータをWWW文書内で特定するための情報を生成するスク

ラップデータ特定情報生成部２０２と、ユーザがデータを指定したWWW文書のＵＲＬとスクラップデータ特定情報の組を記憶するスクラップ情報記憶部２０３と、スクラップページ更新部２０７と、を備え、スクラップページ更新部２０７は、指定されたＵＲＬに対応するWWW文　をWWWサイトから取得するWWW文書取得部２０５と、スクラップデータ特定情報に基づき新たに取得したWWW文書の一部を切り出すデータ抽出部２０４と、抽出したデータを連結し１つの文書にまとめる抽出データ連結部２０６と、を備えている。

【００２３】以下では、ユーザがユーザインタフェース２０１上で指定したデータを「スクラップデータ」、スクラップデータの開始および終了箇所をWWW文書内で特定するためにスクラップデータをスクラップデータ特定情報生成部２０２で生成する情報を「スクラップデータ特定情報」、ユーザがスクラップデータを指定したWWW文書のＵＲＬとスクラップデータ特定情報の組を「スクラップ情報」と呼ぶ。

【００２４】スクラップデータを指定するためのインタフェース（図２の２０１）としては、ユーザが必要とするデータを含むWWW文　のＵＲＬ、およびその文書中の特定データの開始および終了箇所を指定できるものであればよい。

【００２５】その一例として、表示中のテキストの選択機能を有するWWWブラウザを、このインタフェースとして利用可能である。

【００２６】WWWブラウザをスクラップデータ指定のインタフェースとして利用した場合、図３を参照して、ユーザは、WWWブラウザにおいて、特定箇所を選択した後（反転表示された部分が選択箇所を示す；図中ハッチングを施した領域）、選択箇所がユーザが必要とするスクラップデータであることをシステムに指示することになる。なお図３は、WWWブラウザ上でのスクラップデータの選択の一例（画面表示の一例）を示した図である。

【００２７】この指示の後、システムは、WWWブラウザに、WWWブラウザが現在表示しているWWW文書のＵＲＬをスクラップ情報記憶部２０３に記憶する。

【００２８】さらに、WWWブラウザが表示中の文書の元になっているＨＴＭＬ（Hyper Text Markup Language；ハイパーテキストマークアップランゲッジ）形式の文　、およびユーザがスクラップデータとして指定したデータをスクラップデータ特定情報生成部２０２に渡す。

【００２９】スクラップデータ特定情報生成部２０２は、該ＨＴＭＬ文　とスクラップデータを基に、スクラップデータの開始および終了箇所をWWW文　中で特定するためのスクラップデータ特定情報を生成し、スクラップ情報記憶部２０３に記憶する。

【００３０】スクラップデータ特定情報生成部２０２において生成されスクラップ情報記憶部２０３に記憶保持される、このスクラップデータ特定情報は、後に、新たに取得したWWW文書からユーザの必要とする情報を抽出するために、データ抽出部２０４で使用されるものである。したがって、WWWサイトにあるWWW文書が変更された後も、その文書中に残される可能性が高い情報である必要がある。

【００３１】このような情報の一例としては、ＨＴＭＬ文書中のタグの種類や順序といった文書構成に関わるものがある。WWWサイトでは、文書構造（見出し、リストの数や順序など）はそのままで、文書が変更されることが多い。このため、スクラップデータを囲んでいるタグが、その文書内で何番目のものであるかといった情報は有用である。

【００３２】また、他の例としては、スクラップデータの開始行の内容、スクラップデータの開始／終了箇所の直前／直後の行の内容がある。通常、ユーザは、WWW文書内で変更される可能性がある箇所をスクラップデータとして指定するが、WWW文書内で変更される箇所の前後の内容は変更されないことが多い。このため、スクラップデータ直前行、開始行、および直後行の内容は有用である。

【００３３】本実施例では、スクラップデータ直前行、開始行、および直後行の内容をスクラップ情報記憶部２０３に記憶するものとする。

【００３４】図５は、本実施例において、スクラップ情報記憶部２０３に記憶されたデータの一例を示す図である。図５を参照して、ユーザがデータを指定したWWW文書のＵＲＬに対応させて、スクラップデータ直前行、スクラップデータ開始行、スクラップデータ直後行が格納されている。

【００３５】スクラップデータ直前行、開始行、および直後行の内容として、スクラップ情報記憶部２０３に記憶するのは、ブラウザ上に実際に表示されるデータのみとする。

【００３６】すなわち、これらの行中に含まれるテキスト、画像を表示する＜ＩＭＧ＞タグ、水平線（水平罫線；Horizontal Rule）を表示する＜ＨＲ＞タグのみを記憶し、テキストを修飾する＜ＦＯＮＴ＞タグなどは記憶しない。

【００３７】ＨＴＭＬのバージョンによりブラウザ上にデータを表示する効果があるタグの種類は異なるが、本実施例では、＜ＩＭＧ＞タグおよび＜ＨＲ＞タグのみと仮定する。

【００３８】したがって、図４に示すＨＴＭＬ文書をWWWブラウザに表示し、図３に示すように、ユーザがスクラップデータ（図３中反転表示部）を指定した場合、スクラップデータ開始行としては、＜ＦＯＮＴ＞タグおよび＜Ｉ＞タグを削除した"１０／２１　１５：００更新"という文字列のみを、スクラップ情報記憶部２０３に記憶する。

【００３９】また、スクラップデータ直後行には、＜ＨＲ＞タグを記憶する。

【００４０】結果として、スクラップ情報記憶部２０３には、スクラップデータ特定情報として、図５の第３行目に示す情報が記憶される。すなわち、スクラップデータの直前行は、「本日のトップニュース」、スクラップデータ開始行は、「１０／２１　１５：００」、スクラップデータ直後行は＜ＨＲ＞となる。

【００４１】なお、図４に示すＨＴＭＬ文書において、＜Ｈ２＞タグは中見出し、＜ＢＲ＞は改行、＜Ｉ＞タグは斜体（イタリック）、＜ＵＬ＞タグは数字なしの箇条書き、＜ＬＩ＞は箇条書きの項目を、それぞれ指定する。

【００４２】図６は、ユーザから最新WWW情報の取得要求があった場合の処理手順をＰＡＤ（problem analysis diagram；木構造チャート）にて示した図である。説明の便宜上、スクラップ情報記憶部２０３に記憶されたｉ番目のＵＲＬをＳｕｒｌ［ｉ］、スクラップデータ直前行の内容をＳｐｒｖ［ｉ］、スクラップデータ開始行の内容をＳｂｇｎ［ｉ］、スクラップデータ直後行の内容をＳｎｘｔ［ｉ］とする。

【００４３】先ず、WWW文書取得部２０５において、Ｓｕｒｌ［ｉ］に対応する最新のWWW文書ｄｏｃ［ｉ］をＨＴＴＰ（Hyper Text Transfer Protocol）に基づき取得する（ステップ６０２）。

【００４４】次にデータ抽出部２０４において、ｄｏｃ［ｉ］から切り出すデータの開始箇所Ｂ［ｉ］をＳｐｒｖ［ｉ］およびＳｂｇｎ［ｉ］を用いて特定する（ステップ６０３）。

【００４５】Ｂ［ｉ］が特定できた場合（ステップ６０５のＹｅｓ）は、さらにデータ抽出部において、ｄｏｃ［ｉ］から切り出すデータの終了箇所Ｅ［ｉ］をＳｎｘｔ［ｉ］を用いて特定する（ステップ６０６）。

【００４６】Ｂ［ｉ］およびＥ［ｉ］が共に特定できた場合には（ステップ６０７のＹｅｓ）、Ｂ［ｉ］とＥ［ｉ］の間にあるテキストおよびそのテキストを囲む全てのタグをｄｏｃ［ｉ］から抽出しＥＸＴ［ｉ］に代入する（ステップ６０８）。

【００４７】Ｂ［ｉ］およびＥ［ｉ］のいずれかが特定できなかった場合には、抽出データ無しとする。

【００４８】全てのスクラップ情報について上記の処理を行った後、抽出データ連結部２０６において、既に抽出したデータＥＸＴ［ｉ］（１＜＝ｉ＜＝ｎ）を１つのＨＴＭＬ文書にまとめる（ステップ６０９）。

【００４９】図７は、データ抽出部２０４において、抽出するデータの開始箇所Ｂ［ｉ］を特定するための処理手順をＰＡＤにて示したものである。

【００５０】まず、ｄｏｃ［ｉ］において、Ｓｐｒｖ［ｉ］とＳｂｇｎ［ｉ］の文字列が連続している箇所を検索する（ステップ７０１）。ただし、検索は、ｄｏｃ［ｉ］から＜ＩＭＧ＞、＜ＨＲ＞以外のタグを除去したものに対して行う。

【００５１】ｄｏｃ［ｉ］の先頭から文書を走査し、最初にマッチした箇所でＳｂｇｎ［ｉ］文字列の先頭箇所をＢ［ｉ］とする（ステップ７０３）。

【００５２】Ｂ［ｉ］を特定できなかった場合には、Ｓｐｒｖ［ｉ］の文字列だけを用いて上記の検索を行い（ステップ７０４）、最初にマッチした箇所の次の行の先頭をＢ［ｉ］とする（ステップ７０６）。

【００５３】この検索においてもＢ［ｉ］を特定できなかった場合には、さらにＳｂｇｎ［ｉ］の文字列だけを用いて上記の検索を行い（ステップ７０７）、最初にマッチした箇所の直前をＢ［ｉ］とする（ステップ７０８）。

【００５４】ここでもＢ［ｉ］を特定できなければ、最終的にＢ［ｉ］を見つけられなかったものとする（ステップ７０９）。

【００５５】データ抽出部２０４において抽出するデータの終了箇所Ｅ［ｉ］は、上記の手続きで特定したＢ［ｉ］からｄｏｃ［ｉ］文書の末尾までの間で検索する。

【００５６】Ｓｎｘｔ［ｉ］の文字列を用いて検索を行い、最初にマッチした箇所の直前の行の行末をＥ［ｉ］とする。マッチする箇所がない場合は、Ｅ［ｉ］は特定できなかったものとする。

【００５７】上記のデータ抽出手続きによれば、WWWサイトのWWW文書が更新され、スクラップ情報として保持していたスクラップデータ開始行、直前行および直後行に一致する内容が元文　から削除されてしまった場合には、ユーザが意図した箇所を抽出できない。

【００５８】例えば、図５のスクラップ情報記憶部２０３の２番目のＵＲＬの文書が、図８（Ｂ）に示すように更新された場合、スクラップデータ直後行にあたる「神奈川の天気」という文字列は、更新された新しい文書から検索されないため、抽出データを特定できない。この場合、ユーザは、元のWWW文書全体をブラウザに表示し、必要な情報を自ら探す必要がある。

【００５９】しかしながら、抽出データ連結部２０６において、図９に示すように、元文　のＵＲＬにハイパーリンクとして挿入しておくことにより、少ない手間で元文書を表示することが可能となり、実用上問題無い。なお、図９は、本実施例のシステムにより生成されたＨＴＭＬ文書の一例を示す図である（"http://www.aa.co.jp/index.html"から抽出できない旨が提示されている）。

【００６０】

【発明の効果】以上説明したように本発明によれば、ユーザは最初にWWW文書において自分が必要とするデータの開始箇所と終了箇所を指定しておけば、以後はシス

テムが最新のWWW文　を取得し、新たに取得した文書からユーザが必要としていると考えられるデータのみを抽出する。したがって、新たに取得したWWW文書中からユーザ自らが自分の必要とする情報を検索する必要はない。このため、ユーザの情報抽出のための作業を特段に軽減し、利便性を特段に向上するという顕著な効果を奏する。

【００６１】また、本発明によれば、ユーザが必要とする情報が複数のWWW文　に存在する場合でも、システムは、ユーザが必要とするデータを各WWW文書から抽出し、抽出したデータを１つの文書にまとめてユーザに提示するため、ユーザを各WWW文書を１つ１つ開いて内容を閲覧する必要はなく、自分の必要な情報のみを一括して閲覧することが可能である。したがって、本発明によれば、WWWサイトから必要な情報を取得し検索するための作業コストを大幅に軽減することが可能である。

【図面の簡単な説明】

【図１】本発明の一実施例のシステム構成を示す図である。

【図２】本発明の一実施例のシステムを示す図である。

【図３】本発明の一実施例を説明するための図であり、WWWブラウザ上でのスクラップデータの選択の例を示す図である。

【図４】本発明の一実施例を説明するための図であり、ＨＴＭＬ文書の例を示す図である。

【図５】本発明の一実施例を説明するための図であり、スクラップ情報記憶部の内容を示す図である。

【図６】本発明の一実施例を処理フローを説明するための図であり、データ抽出アルゴリズムを表すＰＡＤ図である。

【図７】本発明の一実施例を処理フローを説明するための図であり、データ抽出アルゴリズムを表すＰＡＤ図である。

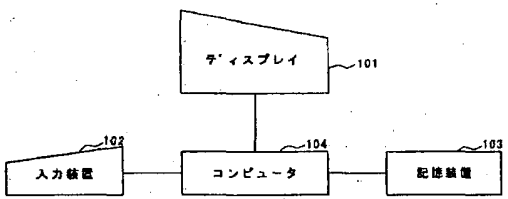
【図８】本発明の一実施例を説明するための図であり、WWW文書の例を示す図である。

【図９】本発明の一実施例を説明するための図であり、システムにより生成されたＨＴＭＬ文書の例を示す図である。

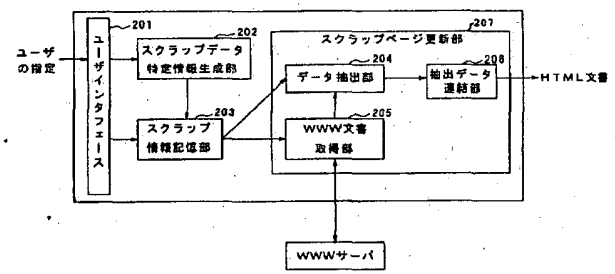
【符号の説明】

１０１　表示装置
１０２　入力装置
１０３　メモリ
１０４　コンピュータ
２０１　ユーザインタフェース
２０２　スクラップデータ特定情報生成部
２０３　スクラップ情報記憶部
２０４　データ抽出部
２０５　WWW文書取得部
２０６　抽出データ連結部

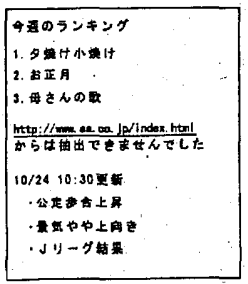
【図１】



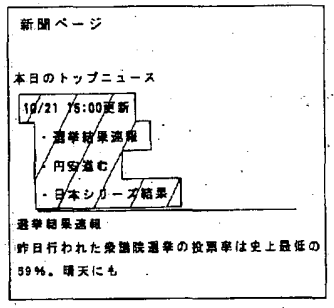
【図２】



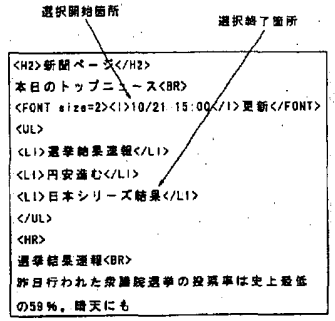
【図９】



【図３】



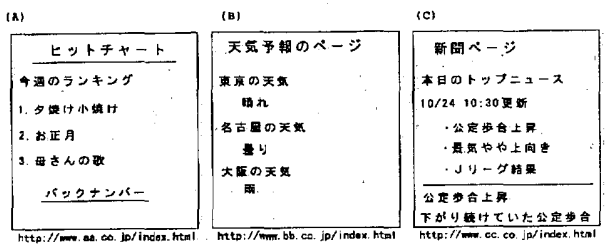
【図４】

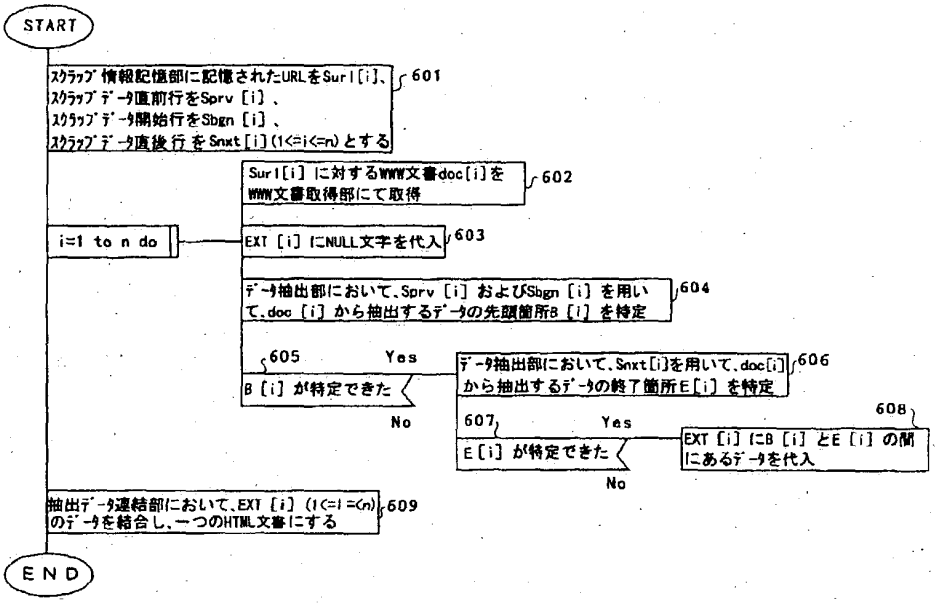


【図５】

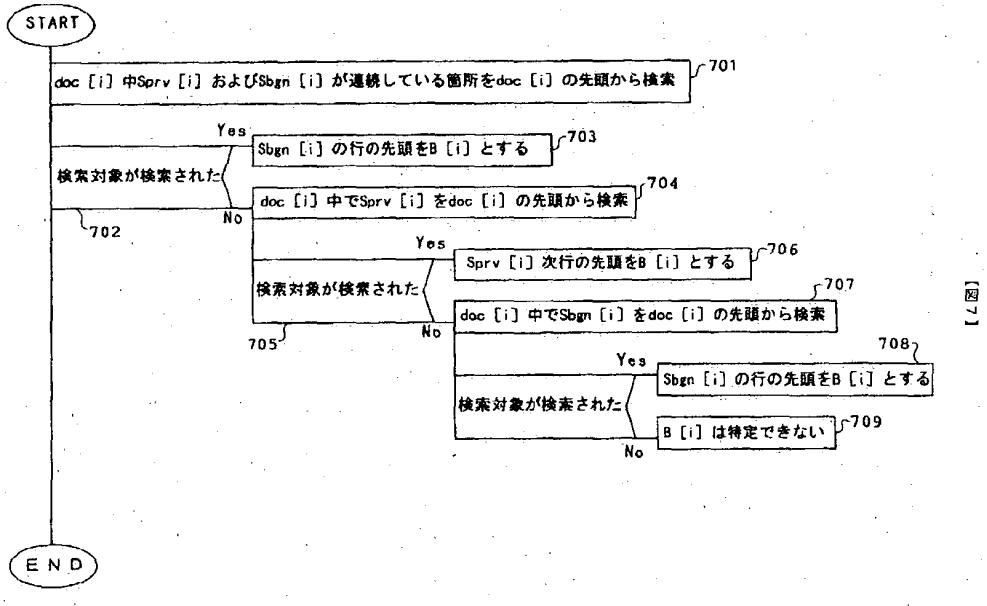
URL	ｽｸﾗｯﾌﾟﾃﾞｰﾀ直前行	ｽｸﾗｯﾌﾟﾃﾞｰﾀ開始行	ｽｸﾗｯﾌﾟﾃﾞｰﾀ直後行
http://www.aa.co.jp/index.html	<HR>	今週のﾗﾝｷﾝｸﾞ	ﾊﾞｯｸﾅﾝﾊﾞｰ
http://www.bb.co.jp/index.html	<IMG SRC=picture.gif>	東京の天気	神奈川の天気
http://www.cc.co.jp/index.html	本日のﾄｯﾌﾟﾆｭｰｽ	10/21 15:00更新	<HR>

【図８】




START
スクラップ情報記憶部に記憶されたURLをSurl[i]、
スクラップデータ直前行をSprv[i]、
スクラップデータ開始行をSbgn[i]、
スクラップデータ直後行をSnxt[i](1<=i<=n)とする
601
Surl[i]に対するWWW文書doc[i]を
WWW文書取得部にて取得
602
i=1 to n do
EXT[i]にNULL文字を代入
603
データ抽出部において、Sprv[i]およびSbgn[i]を用いて、doc[i]から抽出するデータの先頭箇所B[i]を特定
604
605
Yes
B[i]が特定できた
No
データ抽出部において、Snxt[i]を用いて、doc[i]から抽出するデータの終了箇所E[i]を特定
606
607
Yes
E[i]が特定できた
No
608
EXT[i]にB[i]とE[i]の間にあるデータを代入
抽出データ連結部において、EXT[i](1<=i<=n)のデータを結合し、一つのHTML文書にする
609
END

【図6】


START
doc[i]中Sprv[i]およびSbgn[i]が連続している箇所をdoc[i]の先頭から検索
701
検索対象が検索された
Yes
Sbgn[i]の行の先頭をB[i]とする
703
No
702
doc[i]中でSprv[i]をdoc[i]の先頭から検索
704
検索対象が検索された
Yes
Sprv[i]次行の先頭をB[i]とする
706
No
705
doc[i]中でSbgn[i]をdoc[i]の先頭から検索
707
検索対象が検索された
Yes
Sbgn[i]の行の先頭をB[i]とする
708
No
B[i]は特定できない
709
END

【図7】